ECE 6882 Osama Yousuf

HW # 4 Date: _____

Question #1

Monte-Carlo Policy Iteration: (Every-Visit)

Pseudocode:

$Q(s,a) = 0, \pi(s) = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}, \gamma = 0.9$

Episode_length = 5, Returns(s,a) = [] $\forall s, a$

Repeat until (policy unchanged from last iteration)

- Generate an episode using π
- For every pair (s,a) in episode:
  - R ← return following occurrence of s,a
  - Append R to Returns(s,a)
  - $Q(s,a) \leftarrow$ Average(Returns(s,a))
- For each s in the episode
  - $\pi(s) \leftarrow \underset{a}{\text{argmax}}\ Q(s,a)$

- When generating episodes, let the starting state be fixed at A. For exploration, let the agent follow an ε-greedy policy with $\varepsilon = 0.4$.

Pseudocode for generating an episode:

procedure gen_episode (in Ep_length, π, ϵ
out episode)

→ s = A, episode = [ ]

→ while len(episode) < Ep_length

→ $a = \begin{cases} \pi(s) & 1-\epsilon \\ a^1 \text{ or } a^2 \text{ randomly} & \epsilon \end{cases}$

→ take a, observe r, move to s'

→ append (s, a, s', r) to episode

→ s = s'

→ return episode.

Iteration 1:

→ episode = gen_episode(...)

↳ [(0,0,0,0), (0,1,1,4), (1,1,0,-1),
(0,0,0,0), (0,0,0,0)]

My mapping is: A → 0

B → 1

states

a^1 → 0

a^2 → 1

actions

Date: _____

which means:

→ episode'

	S	a	S'	r
1	A	a^1	A	0
2	A	a^2	B	4
3	B	a^2	A	-1
4	A	a^1	A	0
5	A	a^1	A	0

→ (s,a) pairs = $(A, a^1), (A, a^2), (B, a^2), (A, a^1), (A, a^1)$

1) for (A, a^1):

→ $R = 0 + 0.9 \times 4 + 0.9^2 \times -1 + 0 + 0$

→ $\boxed{R = 2.79}$

→ Returns (s, a) = Returns $(A, a^1) = [2.79]$

→ $Q(s, a) = Q(A, a^1) = 2.79$

2) for (A, a^2):

→ $R = 4 + 0.9 \times -1 = 3.1$

→ Returns $(A, a^2) = [3.1]$

→ $Q(A, a^2) = 3.1$

3) for (B, a^2):

→ $R = -1$

→ Returns $(B, a^2) = [-1]$

→ $Q(B, a^2) = -1$

4) for (A, a^1):

→ $R = 0$

→ Returns $(A, a^1) = [2.79, 0]$

→ $Q(A, a^1) = (2.79 + 0)/2 = 1.395$

Date:

5) for (A, a^1):

→ $R = 0$

→ $Returns(A, a^1) = [2.79, 0, 0]$

→ $Q(A, a^1) = (2.79 + 0 + 0)/3$

→ $Q(A, a^1) = 0.93$

All (s, a) pairs visited.

$$Q = \begin{matrix} \\ A \\ B \end{matrix} \begin{matrix} \begin{bmatrix} 0.93 & 3.1 \\ 0 & -1 \end{bmatrix} \\ a^1 \quad a^2 \end{matrix}$$

$$Returns = \begin{matrix} \\ A \\ B \end{matrix} \begin{matrix} \begin{bmatrix} [2.79, 0, 0] & [3.1] \\ [\,] & [-1] \end{bmatrix} \\ a^1 \quad a^2 \end{matrix}$$

$$\pi(s = A) = \underset{a}{\text{argmax}} \underset{a^1}{(0.93} \underset{a^2}{, 3.1)} = a^2$$

$$\pi(s = B) = \underset{a}{\text{argmax}} \underset{a^1}{(0} \underset{a^2}{, -1)} = a^1$$

Thus, $\pi^1 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$

Date: _____

Iteration 2:

→ episode = gen_episode(...)

→ episode^2

	S	a	S'	r
1	A	a^2	B	4
2	B	a'	B	5
3	B	a'	B	5
4	B	a'	B	5
5	B	a'	B	5

1) for (A, a^2):

→ $R = 4 + 0.9 \times 5 + 0.9^2 \times 5 + 0.9^3 \times 5 + 0.9^4 \times 5$

$R = 19.4755$

→ Returns $(A, a^2) = [3.1, 19.4755]$

→ $Q(A, a^2) = (3.1 + 19.4755)/2$

$Q(A, a^2) = 11.28755$

2) for (B, a'):

→ $R = 5 + 0.9 \times 5 + 0.9^2 \times 5 + 0.9^3 \times 5$

$R = 17.195$

→ Returns $(B, a') = [17.195]$

→ $Q(B, a') = 17.195$

3) for (B, a'):

→ $R = 5 + 0.9 \times 5 + 0.9^2 \times 5$

$R = 13.55$

→ Returns $(B, a') = [17.195, 13.55]$

→ $Q(B, a') = (17.195 + 13.55)/2$

$Q(B, a') = 15.3725$

4) for (B, a'):

→ $R = 5 + 0.9 \times 5 = 9.5$

→ $Returns(B, a') = [17.95, 13.55, 9.5]$

→ $Q(B, a') = 13.415$

5) for (B, a'):

→ $R = 5$

→ $Returns(B, a') = [17.95, 13.55, 9.5, 5]$

→ $Q(B, a') = 11.31125$

All (s, a) pairs visited.

$$Q = \begin{matrix} \\ A \\ B \\ \\ \end{matrix} \begin{matrix} \begin{matrix} a^1 & a^2 \end{matrix} \\ \begin{bmatrix} 0.93 & 11.28775 \\ 11.3125 & -1 \end{bmatrix} \end{matrix}$$

$$Returns = \begin{bmatrix} [2.79, 0, 0] & [3.1, 19.4755] \\ [17.95, 13.55, 9.5, 5] & [-1] \end{bmatrix}$$

$\pi(s = A) = \underset{a}{\text{argmax}}\ (0.93, 11.28) = a^2$

$\pi(s = B) = \underset{a}{\text{argmax}}\ (11.31, -1) = a^1$

Thus, $\pi' = \begin{bmatrix} a^2 & A \\ a^1 & B \end{bmatrix}$

Policy π' is the same as π, so we terminate.

Question 2

Tabular Actor-Critic

Pseudocode:

$V(s) = 0$, $H(s,a) = 0$ for all $s \in S$, $a \in A$

Repeat for N episodes:

$t = 0$, $s_0 \in S$ (random)

while $t <$ episode_length:

- $\pi(a|s) = \dfrac{e^{H(s,a)}}{\sum_{a' \in A} e^{H(s,a')}} \quad \forall s, a$
- select action: $a_t \sim \pi(\cdot|s_t)$
- take action a_t, move to s_{t+1}, observe r_{t+1}
- $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$
- $V(s_t) = V(s_t) + \alpha \delta_t$
- $H(s_t, a_t) = H(s_t, a_t) + \beta \delta_t (1 - \pi(a_t|s_t))$
- $t = t + 1$

We are provided with an episode's state-action-reward information, and require computing intermediate policies (π), state values (V), and preferences (H).

Date: ______

$\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.9$

$V(s) = \begin{bmatrix} 0 & 0 \end{bmatrix}$ (columns: A, B)

$H(s,a) =$

	a^1	a^2
A	0	0
B	0	0

t = 0:

→ $s_t = A$, $a_t = a^1$, $r_{t+1} = 10$

$s_{t+1} = A$

~~$\pi(s=A$~~

→ $\pi(a=a^1 | s=A) = \dfrac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \dfrac{e^0}{e^0+e^0}$

$\pi(a^1 | A) = 0.5$

Similarly,

$\pi(a^1|B) = \pi(a^2|A) = \pi(a^2|B) = 0.5$

→ $\delta_t = 10 + 0.9 \times V(A) - V(A) = 10 + 0 - 0$

$\delta_t = 10$

→ $V(A) = V(A) + 0.5 \times 10 = 5$

→ $H(A,a^1) = 0 + 0.1 \times 10 \times (1 - \pi(a^1|A))$

$= 0.1 \times 10 \times (1 - 0.5)$

→ $H(A,a^1) = 0.5$

Date: _____

do,

$$\pi^0 = \begin{array}{c|cc} & a^1 & a^2 \\ A & 0.5 & 0.5 \\ B & 0.5 & 0.5 \end{array}$$

$$V^0 = \begin{bmatrix} A & B \\ 5 & 0 \end{bmatrix}$$

$$H^0 = \begin{array}{c|cc} & a^1 & a^2 \\ A & 0.5 & 0 \\ B & 0 & 0 \end{array}$$

$t = 1$:

→ $s_t = A$, $a_t = a^2$, $r_{t+1} = -5$, $s_{t+1} = B$

→ $\pi(a^1 | A) = e^{H(A,a^1)} / (e^{H(A,a^1)} + e^{H(A,a^2)})$

$= e^{0.5} / (e^{0.5} + e^0)$

⇒ $\boxed{\pi(a^1 | A) = 0.622}$

Similarly,

$\pi(a^2 | A) = e^0 / (e^{0.5} + e^0)$

$\boxed{\pi(a^2 | A) = 0.377}$

$\pi(\cdot | B)$ remain the same.

→ $\delta_t = -5 + 0.9 \times V(B) - V(A)$

$\delta_t = -5 + 0.9 \times 0 - 5 = -10$

→ $V(A) = 5 + 0.5 \times -10 = 5 - 5$

$\boxed{V(A) = 0}$

→ $H(A, a^2) = 0 + 0.1 \times -10 \times (1 - \pi(a^2 | A))$

$= 0.1 \times -10 \times (1 - 0.377)$

$\boxed{H(A, a^2) = -0.623}$

Date: _____

So,

$$\pi' = \begin{bmatrix} 0.622 & 0.377 \\ 0.5 & 0.5 \end{bmatrix}$$

$$V' = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

$$H' = \begin{bmatrix} 0.5 & -0.623 \\ 0 & 0 \end{bmatrix}$$

t=2:

→ $S_t = B$, $a_t = a'$, $\gamma_{t+1} = 40$, $S_{t+1} = A$

→ $\pi(a_1 | A) = e^{0.5} / (e^{0.5} + e^{-0.623})$

$\boxed{\pi(a_1 | A) = 0.7545}$

$\pi(a_2 | A) = e^{-0.623} / (e^{0.5} + e^{-0.623})$

$\boxed{\pi(a_2 | A) = 0.245}$

$\pi(. | B)$ unchanged.

→ $\delta_t = 40 + 0.9 \times V(A) - V(B)$

$\delta_t = 40 + 0.9 \times 0 - 0$

$\delta_t = 40$

→ $V(B) = 0 + 0.5 \times 40$

$\boxed{V(B) = 20}$

→ $H(B, a') = 0 + 0.1 \times 40 \times (1 - 0.5)$

$\boxed{H(B, a') = 2}$

So,

$$\pi = \begin{bmatrix} 0.7545 & 0.245 \\ 0.5 & 0.5 \end{bmatrix}$$

$$V = \begin{bmatrix} 0 & 20 \end{bmatrix}$$

$$H = \begin{bmatrix} 0.5 & -0.623 \\ 2 & 0 \end{bmatrix}$$

Date:

$t=3$:

→ $s_t = A, a_t = a^2, r_{t+1} = -5, s_{t+1} = B$

→ $\pi(a^1|B) = e^2/(e^2+e^0)$

$\boxed{\pi(a^1|B) = 0.88}$

$\pi(a^2|B) = e^0/(e^0+e^2)$

$\boxed{\pi(a^2|B) = 0.12}$

$\pi(\cdot|A)$ unchanged

→ $\delta_t = -5 + 0.9 \times V(B) - V(A)$

$= -5 + 0.9 \times 20 - 0 = 13$

→ $V(A) = 0 + 0.5 \times 13$

$\boxed{V(A) = 6.5}$

→ $H(A, a^2) = -0.623 + 0.1 \times 13 \times (1 - 0.245)$

$\boxed{H(A, a^2) = 0.358}$

so,

$$\pi^3 = \begin{bmatrix} 0.7545 & 0.245 \\ 0.88 & 0.12 \end{bmatrix}$$

$$V^3 = \begin{bmatrix} 6.5 & 20 \end{bmatrix}$$

$$H^3 = \begin{bmatrix} 0.5 & 0.358 \\ 2 & 0 \end{bmatrix}$$

$t=4$:

→ $s_t = B, a_t = a^2, r_{t+1} = 20, s_{t+1} = A$

⇒ $\pi(a^1|A) = e^{0.5}/(e^{0.5} + e^{0.358})$

$\boxed{\pi(a^1|A) = 0.5354}$

⇒ $\pi(a^2|A) = e^{0.358}/(e^{0.5} + e^{0.358})$

$\boxed{\pi(a^2|A) = 0.4646}$

$\pi(\cdot|B)$ unchanged.

Date: _____

→ $\delta_t = 20 + 0.9 \times V(A) - V(B)$

$= 20 + 0.9 \times 6.5 - 20 = 5.85$

→ $V(B) = 20 + 0.5 \times 5.85$

$\boxed{V(B) = 22.925}$

→ $H(B, a^2) = 0 + 0.1 \times 5.85 \times (1 - \pi(a^2|B))$

$= 0.1 \times 5.85 \times (1 - 0.12)$

$\boxed{H(B, a^2) = 0.515}$

So,

$$\pi^4 = \begin{bmatrix} 0.5354 & 0.4646 \\ 0.88 & 0.12 \end{bmatrix}$$

$$V^4 = \begin{bmatrix} 6.5 & 22.925 \end{bmatrix}$$

$$H^4 = \begin{bmatrix} 0.5 & 0.358 \\ 2 & 0.515 \end{bmatrix}$$

$t = 5$:

→ $s_t = A$, $a_t = a^1$, $r_{t+1} = 10$, $s_{t+1} = A$

→ $\pi(a^1|B) = e^2 / (e^2 + e^{0.515})$

$\boxed{\pi(a^1|B) = 0.815}$

$\pi(a^2|B) = e^{0.515} / (e^2 + e^{0.515})$

$\boxed{\pi(a^2|B) = 0.185}$

$\pi(\cdot|A)$ unchanged.

→ $\delta_t = 10 + 0.9 \times V(A) - V(A)$

$= 10 + 0.9 \times 6.5 - 6.5 = 9.35$

→ $V(A) = V(A) + 0.5 \times \delta_t = 6.5 + 0.5 \times 9.35$

$\boxed{V(A) = 11.175}$

→ $H(A, a^1) = 0.5 + 0.1 \times 9.35 \times (1 - \pi(a^1|A))$

$\boxed{H(A, a^1) = 0.9344}$

Date:

So,

$$\pi^5 = \begin{bmatrix} 0.5354 & 0.4646 \\ 0.815 & 0.185 \end{bmatrix}$$

$$V^5 = \begin{bmatrix} 11.175 & 22.925 \end{bmatrix}$$

$$H^5 = \begin{bmatrix} 0.9344 & 0358 \\ 2 & 0.515 \end{bmatrix}$$

The final policy will be,

$$\pi(a^1|A) = e^{0.9344}/(e^{0.9344} + e^{0358})$$
$$= 0.64$$
$$\pi(a^2|A) = 0.36$$
$$\pi(\cdot|B) \text{ unchanged}$$

Thus,

$$\pi^6 = \begin{matrix} \\ A \\ B \end{matrix} \begin{matrix} a^1 & a^2 \end{matrix} \begin{bmatrix} 0.64 & 0.36 \\ 0.815 & 0.185 \end{bmatrix}$$

× ——— × ——— ×